**Optimization**

Basic idea: following up the algorithm in lecture note.

* Calculate the token frequency
* Do prefix filtering, emit (prefix item, item set)
* Group by prefix item, then compute similarity score for each candidate pairs
* Remove duplicate

What I did for improvement:

Before: Use list to store the token order

Now: Use Hashmap to store the sorted tokens

Before: Use “for loop “to find the prefix items for each item set.

Now: Use sortby function combined with hashmap of sorted order when prefix filtering

**The final runtime records (also check the attached runtime.jpg)**

**2 workers time: 38 min 14s = 2294s**

**3 workers time: 32 min 23s = 1943s**

**Chart, line chart

Description automatically generated**